Thesis/
Reports
Teisl

TEST-RETEST RELIABILITY OF DICHOTOMOUS CHOICE CONTINGENT VALUES : AN APPLICATION OF FULL PANEL DESIGN

by
M.F. Teisl, D.W. McCollum,
K.J. Boyle & S.D. Reiling

Test-retest Reliability of Dichotomous Choice Contingent Values: An Application of Full Panel Design[a]

By

Mario F. Teisl
Daniel W. McCollum
Kevin J. Boyle
and
Steven D. Reiling[b]

a This project was financed in part by the State of Maine, Department of Inland Fisheries and Wildlife, Appropriation Account Number 1550.5067, and the Rocky Mountain Forest and Range Experiment Station. Additional funding was provided through the Pittman-Robertson (Wildlife Restoration) and Dingell-Johnson (Fisheries Restoration) Federal Aid Acts, and the Maine Agricultural Experiment Station. The helpful comments on earlier drafts by Nancy Bockstael and Laura Ludwig are greatly appreciated. All errors and omissions, however, are the sole responsibility of the authors.

b The authors are Research Assistant, Department of Agricultural and Resource Economics, University of Maryland; Economist, Rocky Mountain Forest and Range Experiment Station; Associate Professor and Professor, Department of Agricultural and Resource Economics, University of Maine.

28-C9-514 FR

LIBRARY COPY
ROCKY MTN. FOREST & RANGE
EXPERIMENT STATION

Introduction

As contingent valuation (CV) has become more accepted as a means of estimating nonmarket values, the focus of the debate has been on searching out it's limitations. Most of this research has focused on determining the validity of CV estimates but, until recently, there has been little movement in studying the reliability issue. As Reiling et al (1990) notes, this is surprising because reliability and validity are closely interrelated.

Reliability and validity are two methods used by social scientists to evaluate techniques which measure human attributes (Musser et al 1988). A research methodology is valid when it provides estimates which are the desired measure of a theoretical construct. For example, the contingent valuation method is valid if it provides estimates of values for non-marketed commodities that would occur if those commodities were marketed.

A reliable measurement technique minimizes the variance of an estimate due to error. For example, in designing a survey instrument to measure non-market values, a dichotomous choice (DC) format may be more reliable than an open-ended (OE) format to measure willingness-to-pay (WTP). The DC format may be more reliable (i.e. the variance of the estimate has a smaller error component) because the DC format reflects a market setting more familiar to respondents.

It is important to note that estimates with a large variance do not necessarily indicate low reliability. For example, if the

true variance of WTP for a commodity is large in the population being studied, then a reliable technique to estimate WTP will also have a large variance over a representative sample of individuals. As such, reliability and validity are two of the necessary conditions for accuracy in contingent valuation.

One measure of reliability is test-retest reliability. Test-retest reliability is concerned with the intertemporal stability of an estimate. If the determinants of value do not differ significantly between two time periods, then reliable estimators should produce the same estimates of WTP in both time periods. Specifically,

$$WTP_{i,0} = WTP_{i,1}$$

where WTP_i is the maximum willingness-to-pay for individual i measured at two different time periods (0,1). Conversely, if the determinants of value differ between the two time periods, then a reliable measuring technique should produce different estimates of WTP.

$$WTP_{i,0} \neq WTP_{i,1}$$

As Musser et al (1988) points out, confirming test-retest reliability for CV estimates is important because "if CV estimates are not stable over time, their methodological foundation is definitely questionable, and their value for policy evaluation is dubious. In addition, it is important to determine the temporal reliability of CV estimates because the use of these estimates are becoming more common in policy formulation and benefit cost analyses. Given the increasing budgetary

constraints faced by public institutions, many are forced to rely on previous estimates of WTP. Because these estimates are being used over increasingly longer periods of time; it is important to determine the "shelf life" of CV results (Loomis 1989).

Previous Work

The standard methodology used to test for temporal reliability of CV estimates has been the test-retest procedure. In this procedure, a group of individuals are surveyed at two different points in time to determine the stability of their WTP for certain commodities. Previous studies have found evidence of test-retest reliability. However, the studies were limited because, except for Reiling et al (1990) and Epp and Gripp (1992), they resurveyed only one group of individuals over two time periods.

One limitation in using only one group of individuals in the test-retest methodology is if the time period between the two surveys is relatively short. Then the study may exhibit testing bias (Campbell and Stanley, 1963) or "carry-over effect" (Allen and Yen, 1979). In essence, the administration of the first survey may influence the responses to the second survey. Several of the previous studies (Kealy, Dovidio and Rockel 1988; Kealy, Montgomery and Dovidio 1990; Loehman and De 1982) resurveyed their samples after a limited time period.

In the Reiling study two separate groups of individuals were surveyed each in different time periods. By using this

methodology, the Reiling study eliminated testing bias. However, in surveying two independent samples and only looking at the stability of mean WTP, they were not able to establish test-retest reliability. This is because they were not able to test whether WTP responses from the same individuals were the same over the two time periods. As Laughland et al (1991) points out, "mean WTP could be constant over time even if individual WTP varied widely".

The Epp and Gripp study surveyed one group of individuals in two time periods, and they also surveyed a second independent sample of individuals in the second time period. To determine whether there was any significant presence of carry-over effect, they tested whether the responses from the test-retest group in the second time period were significantly different from the responses made by the second independent group. They found no significant carry-over effect. However, their results may not be generalizable because their sample may not be representative due to the low response rate to their survey (39 percent).

There is a second limitation in resurveying only one group of individuals over two time periods. All previous studies have had to make the assumption that the true WTP value in time one is the same as the true value in time two. This is probably not a serious assumption for studies which use a short time period between survey administrations. However, as the time period increases there is an increased chance the determinants of value will have changed. This limitation exists because reliable

estimators should produce different estimates of WTP if the true value differs between two time periods. By lacking a control group, there has been no direct method of testing whether the WTP value had changed during the two time periods.

In order to rectify these limitations, the methodology used in this study incorporates a full panel design (Solomon 1979). In this methodology there are at least four groups or cells of survey participants (Figure 1). Two groups of participants are surveyed in the first time period (groups A1 and B). In the second time period, all of the respondents from group A1 are resurveyed (group A2) and a previously unsurveyed sample of individuals is surveyed (group C). Almost all previous studies have only used the test-retest groups (groups A1 and A2) to test for test-retest reliability of contingent values.[1] By using the control panel design, i.e. by incorporating the control groups B and C, we can directly test whether test-retest reliability exists even if the true value of WTP in time one is different from the true value in time two.

Procedures

In 1989, the State of Maine contracted the Department of Agricultural and Resource Economics at the University of Maine to survey individuals who participated in the 1989 Maine moose hunt. The Maine moose hunt is a six day hunt held approximately the

[1] Reiling et al used groups A1 and C; Epp and Gripp used groups A1, A2 and C.

first week in October. There is an annual lottery to determine the 1,000 hunters (900 residents and 100 nonresidents) that will be legally allowed to hunt. On average, over 65,000 Maine residents participate in the moose hunt lottery.

To develop our samples of moose hunters, we conducted a mail survey of all licensed moose hunters in Maine during 1989. We administered the survey by following the Total Design Method, outlined by Dillman, because it is designed to obtain a high survey response rate and a sample data set which is representative of the population being surveyed (Dillman, 1978).

We incorporated a full-panel design into the administration of the survey so that we could test the test-retest reliability of dichotomous choice contingent values (DC-CV). In late October 1989, we surveyed 550 of the 900 residents who obtained a 1989 Maine resident moose hunting permit (Figure 1). This sample was split into two groups; 150 respondents were placed in group A1 and the remaining 400 were placed in group B. The response rate to this survey was 97 percent for group A1 and 93 percent for group B.

During March 1990, we resurveyed the members of group A1 and designated these respondents as group A2. At this time we also surveyed an additional 150 individuals (group C) who participated in the 1989 Maine moose hunt but had previously been unsurveyed. The response rates to these surveys were 82 percent for the retest (A2) group and 92 percent for the new respondents. In reporting our results of the test-retest group (A1-A2), we are

restricted to those individuals who returned usable surveys in both time periods.

To elicit accurate DC-CV estimates we began the questionnaire with several questions relating to the respondents general hunting and general moose hunting experience. Next, the questionnaire focused on the respondent's moose hunting experience during the 1989 moose hunt. To make the individual's budget constraint salient, we then asked respondents to fill out an expense table for their 1989 moose hunt. Specifically, we asked them to estimate the amount they spent for 12 expenditure categories and then to provide a total amount spent on their 1989 moose hunt.

Immediately following the expenses table, we asked respondents the DC-CV question presented in Figure 2. The underlined blank in the DC-CV question was filled-in with bids generated from a cumulative logit distribution. The distribution from which the bid estimates were drawn was based upon responses to an open-ended CV question in a moose hunting survey we administered following the 1988 Maine moose hunt. The bid levels used in the 1989 survey were randomly assigned to respondents in groups A1, B and C. However, the individuals in group A2 were presented the same bids that they were presented when they were treated as group A1.

Due to the nature of this study, we needed to determine that the control group (groups B and C) for each time period had the same characteristics as the test-retest group (groups A1 and A2).

In turn, we first analyzed the data by using a multivariate analysis of variance (MANOVA) technique. MANOVA compares the differences between the means of several variables across groups of subjects. MANOVA is appropriate when testing between-group differences on a set of variables (Hand and Taylor, 1987). A benefit of using MANOVA over multiple t-tests or multiple univariate ANOVA tests on a number of variables is that it decreases the probability of incorrectly rejecting the null hypothesis of no difference between the groups (a Type I error).

To determine the test-retest reliability of DC-CV estimates we used the following methodologies. First, we used correlation analysis on the responses of the test-retest group (groups A1 and A2). Correlation analysis gives us a direct measure of the degree to which responses from the same individuals are the same in two time periods. Following Musser et al (1988), we developed a 2 X 2 contingency table and used a chi-square test and calculated a Cramer's coefficient.[2]

We cannot develop a contingency table for the control group because groups B and C are composed of different individuals. Therefore, we can no longer directly measure whether the responses in the two time periods are the same. However, if the MANOVA analysis indicates the four groups have similar characteristics (i.e. are drawn from the same distribution), then

---

[2] The Cramer's coefficient measures the degree of association between responses from the same individuals in the two time periods.

we can use indirect measures of association. In turn, we use chi-square tables to determine if the distributions of the responses to the DC-CV question are similar (i.e. drawn from the same population) between the four groups.

Although the above methods can be used to test for temporal reliability of the responses to DC-CV questions, a more compelling question is whether the estimated values for mean WTP are reliable. Specifically, we want to test whether

$$E\{WTP_{t=0}\} = E\{WTP_{t=1}\}$$

where E{} is the expectations operator.

We began this stage of our analysis by first estimating separate logit equations for each of the four groups separately (discussion of the model specification is deferred until the next section). Then, we pooled the groups and re-estimated the logit equation. By using a likelihood ratio test, we test whether the coefficients of the WTP functions are stable (i.e. whether the responses for each of the four groups are drawn from the same population).

A final test of test-retest reliability involves estimating the mean WTPs, and constructing confidence intervals of the WTP estimates and observing the degree of overlap between the different groups. As Park et al (1991) correctly points out, without confidence intervals it is difficult to determine whether WTP measures are statistically different.

There are four principle methods that have been used to construct confidence intervals around DC-CV welfare measures; the

bootstrap, the jackknife, Cameron's linearization, and the Krinsky and Robb technique (Cooper 1993). Kling (1991) and Cooper (1993) used Monte Carlo simulation to examine the accuracy of these methods.

Kling, using linear functions, compared the Cameron, the bootstrap and the Krinsky and Robb techniques and found all three approaches provide reasonable approximations of the true confidence intervals. Cooper (1993) compared all four approaches, using logit functional forms, and found they all produced similar upper and lower bounds to the confidence intervals. However, Cooper found the Cameron, and the Krinsky and Robb approaches have a tendency to underpredict the size of the confidence intervals.[3]

Given the purpose of this paper, we did not want to incorrectly accept the null hypothesis of no difference in mean WTP. We felt that using the Cameron, and Krinsky and Robb approaches, and a 90 percent confidence interval, would be conservative. Further discussion of the confidence interval estimation is deferred until after the next section.

Theoretical and Empirical Model

Given that DC-CV is used to obtain estimates of consumer surplus for nonmarketed commodities, it is important the empirical model is based upon a model of utility maximizing

[3] For a detailed discussion of how the confidence intervals are estimated see Cameron (1988 and 1991) and Park, Loomis and Creel (1991).

behavior. Hanemann (1984), using a utility difference model, provides a methodology which derives Hicksian compensating or equivalent variation measures. The model explicitly recognizes the utility maximization choice underlying an individual's responses to DC-CV questions.

In developing the model, we assume hunters derive utility from hunting and income. Furthermore, the hunter's utility function is assumed to be an increasing function with respect to the quality of the hunt. The individual's level of utility can be represented by either $u_i^0 = (0, y_i, q_i, s_i)$, if the individual does not hunt, or by $u_i^1 = (1, y_i - c, s_i)$, if the individual hunts. Hunting is denoted by 1, not hunting is denoted by 0; income is denoted by y_i; c denotes the cost of hunting; q_i denotes the quality of the hunt experienced; other individual characteristics assumed to affect preferences are denoted by s_i; the subscript 'i' indexes the individual.

We assume the individual knows with certainty their utility function u_i. However, from the researchers point of view, the utility function contains unobservable elements such as the characteristics of the hunter or characteristics of the choice set available to the hunter. In assuming the unobservable elements are stocastic, we can rewrite the hunter's utility as

$u_i^1 = v(1, y_i - c_i, q_i, s_i) + e_{1i}$ if they hunt or

$u_i^0 = v(0, y_i, s_i) + e_{0i}$ if they do not hunt.

The e_i are assumed identically and independently distributed with mean zero.

In modeling the hunters' decision, a hunter will hunt if

$$u_i^1 > u_i^0,$$

or equivalently if

$$v(1, y_i - c_i, q_i, s_i) + e_{1i} > v(0, y_i, s_i) + e_{0i}.$$

When asking a hunter a DC-CV question, researchers are confronted with a binary variable, either the individual will hunt or they will not. Given the individual's response is based upon their utility function, which is hidden to the researcher, it is a random variable whose probability distribution is

$$\text{Prob(YES)} = \Pr\{v(1, y_i - c_i, q_i, s_i) + e_{1i} > v(0, y_i, s_i) + e_{0i}\}$$

or in utility differences

$$\text{Prob(YES)} = \Pr\{v(1, y_i - c_i, q_i, s_i) - v(0, y_i, s_i) > e_{0i} - e_{1i}\}$$

and

$$P0 = \text{Prob(NO)} = 1 - \text{Prob(YES)}$$

If we define

$$\Theta = e_{0i} - e_{1i}$$

and

$$\Delta v = v(1, y_i - c_i, q_i, s_i) - v(0, y_i, s_i)$$

then

$$\text{Prob(YES)} = F_\Theta(\Delta v) \text{ is the CDF of } \Theta$$

where Δv represents the probability of a 'yes' response to the DC-CV, and is a function of the bid amount the individual is asked to pay. When $\Delta v > 0$ (i.e. when the increase in utility by having access to the resource is greater than the decrease in utility caused by the decline in income) then respondents will maximize utility by answering 'yes' to the DC-CV question.

Following Loomis et al (1991), the indirect utility difference model yields the logit specification when the probability of a 'yes' response is specified as the CDF of a standard logistic variate:

$$P(\text{yes}) = [1+\exp(-\Delta v)]-1 \qquad (1)$$

The linear logit model is one of the model specifications for Δv, outlined by Hanemann (1984), compatible with the utility maximization assumption. The linear logit model uses a simple form of the utility function

$$v_0 = \alpha_0 + \beta_1 y + \beta_2 q + \beta_3 s + e_0$$

$$v_1 = \alpha_1 + \beta_1 (y-c) + \beta_3 s + e_1$$

or in utility differences

$$\Delta v = (\alpha_1-\alpha_0) - \beta_1 c + \beta_2 q + \Theta$$

Note that in taking differences, the variables denoting income and the individual's characteristics drops out of the model. Thus, this utility function implies no income effects making the Hicksian compensating and equivalent variations both equal to 'c'. Thus, 'c' is the price that makes the hunter indifferent between hunting or not hunting.

The estimated model took the form of (1) where

$$\Delta v = \tau_0 - \tau_1 c + \tau_2 q_1 + \tau_3 q_2 + \Theta, \qquad (2)$$

$\tau_1 = (\alpha_1-\alpha_0)$ is the intercept, 'c' is the bid provided in the DC-CV question, q_1 and q_2 are two qualitative variables to denote the quality of the hunt. Specifically, when $q_1 = 1$ and $q_2 = 0$, then the hunter bagged a cow moose; when $q_1 = 0$ and $q_2 = 1$, then

the hunter bagged a bull moose; when $q_1 = q_2 = 0$ then the hunter did not bag a moose.[4] Θ is as defined previously.

Finally, Hanemann shows that expected WTP is computed as

$$E\{WTP_{ti}\} = \tau^{*} / (-\tau_1) \qquad (3)$$

where, for our purposes, $\tau^{*} = (_{0i} + \tau_2 q_{1i} + \tau_3 q_{2i})$, and $-\tau_1$ is the coefficent on the bid variable, both from the estimated model.

Confidence Interval Estimation

The Cameron approach to estimating expected WTP uses a censored logistic regression model. Specifically, Cameron assumes

$$WTP_i = x_i\beta + u_i \qquad (4)$$

where WTP_i is the true, unobserved WTP for individual 'i', x_i is a vector of explanatory variables, u_i is assumed identically and independently distributed with mean zero. McConnell (1990) proved Cameron's approach is the dual to the Hanemann linear utility difference model when the bid is additively separable. Where Hanemann's model is based on the difference in utility functions, Cameron's is based on the difference in cost functions.

Given Cameron's approach estimates expected WTP as a linear, monotonic transformation of (3), the confidence intervals can be

---

[4] The original model specification included only one qualitative variable to denote whether the hunter bagged a moose. However, over 96 percent of the moose hunters bagged a moose. As a result there was multicollinearity between this variable and the constant term. In turn, the model was respecified to reduce the level of multicollinearity.

estimated using an approach similar to OLS. The confidence interval is estimated as

$$CI[E\{WTP_{ti}\}] = x\beta \pm t_{\alpha} (x\Sigma_{\beta}x)^{0.5}$$

where $x\beta$ is from (4), t_{α} is the appropriate statistic from a t distribution, and Σ_{β} is the estimated covariance matrix from (4).

In contrast, the Krinsky and Robb (KR) technique uses a simulation method directly from the Hanemann model. Bockstael and Strand (1987) have shown that parameter estimates, such as the τ vector in (2), are random variables. The KR technique uses the variability in the τ vector to estimate a confidence interval around the mean WTP. The KR technique can be used to estimate the distribution of estimators, which are a non-linear function of estimated parameters (Loomis 1991). Equation (3) is just such an equation.

The KR technique uses multiple random drawings of the estimated τ vector from (2) based on a multivariate normal distribution with mean τ and variance -covariance Σ_{τ}.[5] Expected WTP is calculated for each drawing of the τ vector and an empirical distribution is constructed from the set of WTP values. A $(1-\alpha)$ confidence interval is then calculated by sorting the WTP values in ascending order and deleting the $\alpha/2$ at each tail (Cooper 1993). Based on Cooper, we set the number of drawings at 1,000.

---

[5] The logit model is asymptotically normal with desirable asymptotic properties (Amemiya 1981).

Results

The results of a MANOVA analysis are quite sensitive to the presence of outliers in the data. This sensitivity exists because the greater the variation in the data, the harder it is for a MANOVA technique to make distinctions between the different groups. Therefore, before performing MANOVA analysis, the frequency distributions of the data were analyzed for outliers or unreasonable responses. Fortunately, no such adjustments needed to be made.

There are four test statistics available for MANOVA analysis; the Pillai-Bartlett trace, Wilk's lambda, the Hotelling-Lawley trace and Roy's largest eigen value statistic. The choice of which test statistic to use is dependent upon the degree to which the data violates the basic MANOVA assumptions (normality, homogeneity of variances and independence), and the absence of equal cell sizes (Hand and Taylor, 1987). Currently, there is no generally accepted method of determining which test statistic to use. For clarity, we decided only to report the Pillai-Bartlett trace because all of the test statistics provided the same result.

The MANOVA analysis found no significant differences in the socioeconomic characteristics between respondents in group A1 and group B (Pillai = .00678, $F_{(4,435)}$ = .99246, p = .396) (Table 1). Likewise, there are no significant differences in the level of participation in outdoor activities between these two groups (Pillai = .00616, $F_{(4,476)}$ = .73761, p = .567). Similarly, we

found no significant differences in the socioeconomic characteristics (Pillai = .00178, $F_{(4,223)}$ = .13336, p = .940) or participation in outdoor activities (Pillai = .02004, $F_{(4,247)}$ = 1.26270, p = .285) between respondents in group A2 and C (Table 2). Therefore, we conclude the sampling strategy for the survey resulted in similar samples of the population of moose hunters in Maine in both time periods. These preliminary tests provides a baseline perspective on whether our control group is the same as our test-retest group during both time periods.

One method of testing test-retest reliability is to analyze a contingency table of responses from the same group of individuals in two time periods (Table 3). In the test-retest group (A1-A2), 35 individuals responded 'yes' and 56 individuals responded 'no' to the DC-CV question in both time periods. Overall, 79 percent of the responses made by these individuals were the same in both time periods. Moreover, there is a significant relationship between the responses in time one and time two for the test-retest group (χ^2 = 39.65, df = 1, p < .01). Finally, the Cramer's V statistic[6] of 0.57, which is significant at the .05 level, indicates there is a correlation between responses in the two time periods.

The contingency table described in the previous paragraph indicates that responses to the DC-CV question seem to be the same in both time periods for the test-retest group. This alone

---

[6] Cramer's V statistic is bounded by zero (indicating no correlation) and one (indicating perfect correlation).

does not provide evidence of test-retest reliability because the determinants of value may differ between the two time periods (previous studies had to assume the true value did not change over the two time periods). In turn, we need to test whether the responses given by the control group (B-C) are the same in the two time periods.

We cannot create a contingency table for the control group similar to the one we created for responses from the test-retest group because the control groups (B and C) are composed of different individuals. However, we can test whether the distribution of DC-CV responses are the same for the control and test-retest groups in both time periods (Table 4).

The percent of individuals in each group that responded 'yes' to the DC-CV question varied from 38 (group A2) to 48 percent (group C). The distributions of responses is not significantly different between the four groups ($\chi^2 = 2.8$, df = 3, $p = 0.43$). This indicates that the distributions of responses made by individuals in the test-retest group and the control group are the same in both time periods. This test, along with the contingency table results indicate that DC-CV estimates are temporally reliable.

Results of the logit analysis also indicate that the four groups are quite similar (Table 5). The chi-square statistics are significant in all five equations, indicating the independent variables are significantly predicting the probability of a 'yes' response. The parameter estimates for the BID variable are

largely uniform between equations and are significant and negative in all five equations at the five level. Thus, higher bid amounts reduce the probability of a 'yes' response.

The quality of hunt variables have positive signs in the group A1, B and pooled equations, and has negative signs in the A2 and C equations. The quality of hunt variables are significant at the five percent level in the group B equation, and at the 10 percent level in the pooled equation. Note the presence of multicollinearity between these variables and the intercept may have reduced the efficiency of these variables.

Given that bagging a bull moose is highly desired by hunters, we hypothesized the BULL variable would have a positive sign. The sign of the COW variable was indeterminate for two reasons. On one hand, bagging a moose should increase the hunters utility, indicating a positive sign on the COW variable. However, given only one moose may be taken, and the high probability the hunter may not get another chance to hunt moose in Maine, there may be a disutility to bagging a cow.[7] Admittedly, this second point is less plausible.

The likelihood ratio test indicates the unrestricted logit equations are not significantly different from the restricted (pooled) equation ($\chi^2 = 5.86$, df = 3, $p < .00$). The above results indicate the distributions of the responses are not significantly different between the test-retest group and the

[7] Less than two percent of the applicants to the Maine moose hunt actually obtain a moose hunting permit. Thus, hunters may view bagging a bull moose as a once in a lifetime opportunity.

control group in either of the two time periods. In other words, the responses are drawn from the same distribution, and thus, have all the same moments. This test strongly indicates DC-CV estimates are temporally reliable.

The estimated mean WTP's are relatively close to each other and the bounds provided by the Cameron and KR approaches are quite similar (Table 6). Mean WTP is lowest for group A2 ($640) and highest for group C ($1001). As expected, the size of the confidence interval decreases with increases in the sample size.

In the test-retest group, the mean WTP decreased in the second time period, whereas, for the control group the mean WTP increased in the second time period. By only looking at the means, this result would seem to indicate DC-CV estimates of WTP are not temporally reliable. However, the estimated confidence intervals have a significant amount of overlap indicating the mean WTP's are the same between the four groups. This provides further evidence that the DC-CV method produces estimates of mean WTP that are temporally reliable.

Conclusions

It is important to determine the test-retest reliability of contingent valuation because its use is becoming more common in policy analyses. Also reliability and validity are two of the necessary conditions for accuracy. The previous methodology used to test for temporal reliability, the test-retest procedure, is somewhat limited because it requires the assumption that the true

WTP value in time one is the same as the true value in time two. Given that reliable estimators should produce different estimates of WTP when the true value differs between two time periods the traditional test-retest procedure is unsatisfying. In order to rectify these limitations, the methodology used here extends the traditional methodology by incorporating a full panel design.

Our results indicate that DC-CV estimates are temporally reliable. By analyzing a contingency table of responses from the same group of individuals in two time periods, we found that 79 percent of the individuals made the same responses in both time periods. Moreover, by using a chi-square table and logit analysis, we were able to determine that the distributions of responses made by individuals in the test-retest group and the control group are the same in both time periods. Finally, estimated confidence intervals indicated that the estimates of mean WTP were the same for all four groups. These tests strongly suggest that dichotomous choice contingent valuation estimates are temporally reliable.

Figure 1. Complete panel design used to test temporal reliability of contingent values.

	Time of survey administration	
	November, 1989	March, 1990
Test-retest group	**Group A1**	**Group A2**
Control group	**Group B**	**Group C**

Note: Groups A1 and A2 are composed of the same individuals; Groups B and C are composed of different individuals.

Figure 2. Dichotomous choice contingent valuation question used in the 1989 Maine moose hunting survey.

Hunting expenses often go up or down. For example, gas prices rose substantially in the 1970's, fell somewhat in the early 1980's and have recently risen again. Would you still have gone moose hunting in Maine during 1989 if your total expenses had been $ _____ more than the total you just calculated?

(CIRCLE ONE NUMBER)

1 YES
2 NO

Table 1. Characteristics of test-retest and control group samples.

	Time Period 1	
	Group A1	Group B
Average age	38	40
Percent male	86 %	81 %
Average education	High school	High school
Average 1988 household income	$ 28,000	$ 29,400
Participation in outdoor activities in Maine during 1989:		
Open water fish	75 %	80 %
Ice fish	64	63
Marine fish	33	32
Hunt (other than moose)	92	91

Table 2. Characteristics of test-retest and control group samples.

	Time Period 2	
	Group A2	Group C
Average age	38	41
Percent male	84 %	81 %
Average education	High school	High school
Average 1988 household income	$ 29,900	$ 29,900
Participation in outdoor activities in Maine during 1989:		
Open water fish	79 %	83 %
Ice fish	59	65
Marine fish	39	32
Hunt (other than moose)	93	92

Table 3. Contingency table of dichotomous choice contingent valuation responses from individuals in the test-retest group (A1 and A2).

Response in November, 1989 (Group A1)	Response in March, 1990 (Group A2)	
	YES	NO
YES	35	15
NO	9	56

Table 4. Chi-square tables showing distributions of dichotomous choice contingent valuation responses from individuals in the test-retest and control groups.

	Test-retest group		Control group	
	Group A1	Group A2	Group B	Group C
Percent Yes	44	38	42	48
Percent No	56	62	58	52

Table 5. Results of unrestricted and restricted logit estimation.

Coefficients and Statistics	Group A1	Group A2	Group B	Group C	Pooled
Intercept	0.4401	1.8789	-0.6131	2.0122	0.4158
	(0.9456)	(0.8651)	(0.5811)	(0.9989)	(0.3518)
Bid	-0.0018	-0.0014	-0.0013	-0.0012	-0.0013
	(0.0004)	(0.0004)	(0.0002)	(0.0003)	(0.0001)
Cow	1.2964	-0.7984	1.3410	-1.1527	0.4939
	(0.9724)	(0.8317)	(0.6047)	(0.9918)	(0.3628)
Bull	1.1361	-1.1697	1.7768	-0.7289	0.7125
	(0.9360)	(0.7940)	(0.5909)	(0.9657)	(0.3495)
χ^2	31.133	20.538	62.028	21.444	117.489
Maximum of Log-Likelihood	-62.326	-65.276	-208.713	-81.372	-427.963

Table 6. Expected value of willingness-to-pay and 90 percent confidence intervals.

	Group A1	Group A2	Group B	Group C	Pooled
Lower bound					
Cameron	$ 641.63	$ 370.89	$ 533.97	$ 735.85	$ 654.78
Krinsky & Robb	610.83	280.50	513.34	709.50	647.58
Mean	$ 839.66	$ 640.54	$ 689.96	$1001.05	$ 760.13
Upper bound					
Cameron	$1037.69	$ 910.18	$ 845.94	$1266.25	$ 865.48
Krinsky & Robb	1041.34	900.49	851.15	1298.76	863.79

References

Allen, Mary J. and Wendy M. Yen 1979. Introduction to Measurement Theory Brooks Cole Publishing.

Amemiya, T. 1981. "Qualitative Response Models" Journal of Economic Literature 19.

Bockstael, Nancy E. and Ivar E. Strand 1987. "The Effects of Common Sources of Regression Error on Benefit Estimates" Land Economics 63.

Campbell, Donald T. and Julian C. Stanley 1963. Experimental and Quasi-Experimental Designs for Research Rand McNally & Co.

Cameron, T.A. 1988. "A New Paradigm for Valuing Non-Market Goods using Referendum Data: Maximum Likelihood Estimation by Censored Logistic Regression" Journal of Environmental Economics and Management 15.

Cameron, T.A. 1991. "Interval Estimates of Non-market Resource Values from Referendum Contingent Valuation Surveys" Land Economics 67(4).

Cooper, Joseph C. 1993. "A Comparison of Approaches to Calculating Confidence Intervals for Benefit Measures From Dichotomous Choice Contingent Valuation Surveys" Economic Research Service, U.S.D.A.

Dillman, Don 1978. Mail and Telephone Surveys John Wiley & Sons.

Efron, B. 1981. "Nonparametric Standard Errors and Confidence Intervals" Canadian Journal of Statistics 9.

Epp, Donald J. and Sharon I. Gripp 1992. "Test-Retest Reliability of Contingent Valuation Estimates for an Unfamiliar Policy Choice: Valuation of Tropical Rain Forest Preservation" Agricultural Economics and Rural Sociology Department Staff Paper 221, Pennsylvania State University.

Hand D.J. and C.C. Taylor Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioral Scientists Chapman and Hall. London 1987.

Hanemann, Micheal W. 1984. "Welfare Evaluations in Contingent Valuation Experiments with Discrete Responses" American Journal of Agricultural Economics 66.

Judge, George C., R. Carter Hill, William E. Griffiths, Helmut Lutkepohl and Tsoung-Chao Lee 1988. Introduction to the Theory and Practice of Econometrics John Wiley & Sons.

Kealy, Mary J., John F. Dovidio and Mark L. Rockel 1988. "Accuracy in Valuation is a Matter of Degree" Land Economics 64(2).

Kealy, Mary J., Mark Montgomery and John F. Dovidio 1988. "Reliability and Predictive Validity of Contingent Valuation: Does the Nature of the Good Matter?" Journal of Environmental Economics and Management 19.

Kling, Catherine L. 1991. "Estimating the Precision of Welfare Measures" Journal of Environmental Economics and Management 21.

Kling, Catherine L. and Richard J. Sexton 1990. "Bootstrapping in Applied Welfare Analysis" American Journal of Agricultural Economics 72(2).

Krinsky, I. and A. L. Robb 1986. "On Approximating the Statistical Properties of Elasticities" Review of Economics and Statistics 68.

Laughland, Andrew, Wesley N. Musser and Lynn M. Musser 1991. "An Experiment on the Reliability of Contingent Valuation" Agricultural Economics and Rural Sociology Department Staff Paper 202, University of Pennsylvania.

Loehman, Edna T. and Vo Hu De 1982. "Application of Stocastic Choice Modeling to Policy Analysis of Public Goods: A Case Study of Air Quality Improvements" Review of Economics and Statistics 64.

Loomis, John B. 1989. "Test-Retest Reliability of the Contingent Valuation Method: A Comparison of General Population and Visitor Responses" American Journal of Agricultural Economics 71(1).

Loomis, John B. 1990. "Comparative Reliability of the Dichotomous Choice and Open-ended Contingent Valuation Techniques" Journal of Environmental Economics and Management 18.

McConnell, K. 1990. "Models for Referendum Data: The Structure of Discrete Choice Models for Contingent Valuation" Journal of Environmental Economics and Management 18.

Mitchell, Robert C. and Richard Carson 1989. Using Surveys to Value Public Goods: The Contingent Valuation Method Resources for the Future. Washington D.C.

Musser, Wesley N., Keith A. Lampi, Lynn M. Musser and Fredrick W. Obermiller 1988. "Test-Retest Reliability of Contingent Valuation Methods" Selected Paper at the annual meeting of the Northeast Agricultural and Resource Economics Association.

Park, Timothy, John B. Loomis and Micheal Creel 1991. "Confidence Intervals for Evaluating Benefits Estimates from Dichotomous Choice Contingent Valuation Studies" Land Economics 67(1).

Reiling, Steven D., Kevin J. Boyle, Marcia L. Phillips and Mark W. Anderson 1990. "Temporal Reliability of Contingent Values" Land Economics 66(2).

Solomon, Richard L. 1979. "An Extension of Control Group Design" Psychology Bulletin 46.